Worldwide Natural Disasters

Analysis of human and economical impact

By Luis Esquivel
Data retrieved March 9, 2023

Abstract

With a Worldwide Natural Disasters (1900-2022) dataset, questions like: what is the deadliest disaster of all?, how often does disasters happen?, could deaths be predicted?, where addressed. Grouping data to report deaths, line graphs to examine trends, and Regression Analysis for deaths prediction were the methods used. It was found that droughts are the the deadliest disasters, floods the most frequent, and the error associated with predicting deaths with this dataset is large.

Motivation

Natural Disasters have impacted mankind since the beginning of times. In recent years, as the world population increases and more people live in highly populated areas, like cities, or are forced to live in dangerous areas, like river margins, the impacts of natural disasters greatly increases.

Studying Natural Disasters and their consecuences can help us prepare so we can minimize their impacts (for example, deaths, homeless, economic losses).

This is important to all people living in areas prone to natural disasters like floods, droughts, storms, earthquakes, landslides, tornados (to mention some). But specially, to authorities responsible of emergencies prevention and response.

Dataset

Source: Centre for Research on the Epidemiology of Disasters (CRED), EM-DAT

Contents and time frame: All Natural Disasters since 1900 to the present, with information on location, date, type of disaster, human and economic losses.

Size: Almost 15,000 records characterized by 50 variables.

Acquisition of dataset: Through the EM-DAT query tool (https://public.emdat.be)

Data Preparation and Cleaning

1. Feature selection: from the 50 features, a total of 18 were selected, considering what was needed to answer the research questions.

2. Handle NaN values: for numerical features like deaths, affected or economic loss, instead of zero, the dataset had blank fields, which pandas interpreted as NaN values. This were replaced by the value 0.0 . For categorical features with NaN values, see section 3.1 of Jupyter Note Book (jnb).

3. A problem arise when ploting the number of disasters over time, the line plot was strange, the line cross itself. It turn out the date were not in cronological order, so the solution was to sort the dataset by date, and problem solved.

Research Questions

1. Which are the natural disasters that produce the greatest human and economic losses?
2. How often does natural disasters happen since 1900?
3. With this data set, could deaths due to natural disasters be predicted, based on location and disaster type?

Methods

1. To identify single disaster events with greatest impact, **maximum values** for losses had to be determined, along with the corresponding event. To identify the type of disaster with greatest impact, the **data grouped by disaster type** was ploted on a **grouped bar plot**, for ease of comparison.

2. Because how often disasters occur over time is frequency, a **line graph** was chosen to explore this. The **data grouped by year** was ploted on a line graph, enabling the identification of spikes or trends.

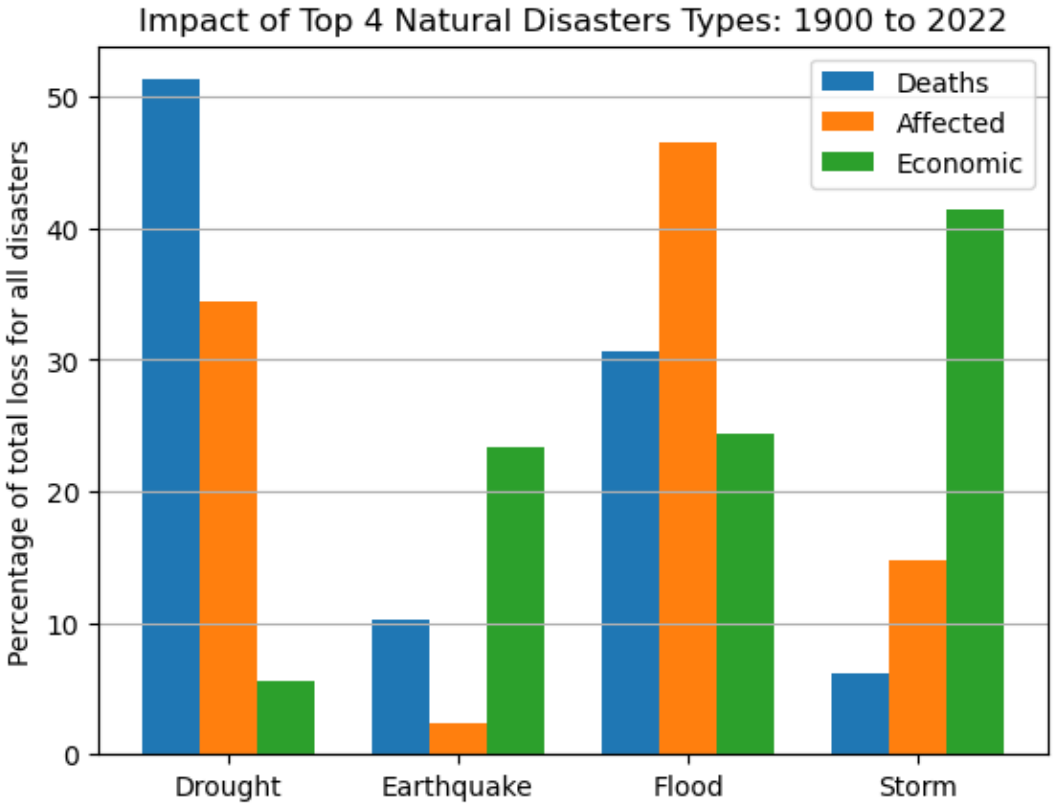
3. Because deaths was the target feature to be predicted, and it is of numerical type, **regression analysis** was used. A linear reggressor and a decision tree regressor were explored.

Findings: Greatest Natural Disasters of All Time?

Impact	Country	date	Disaster Type	Total Deaths	Total Affected	Total Damages, Adjusted ('million US$)
Deadliest	China	1931-07-01	Flood	3,700,000	0	26,881
Affecting more people	India	2015-01-01	Drought	0	330,000,000	3,704
Greatest ecocomic impact	Japan	2011-03-11	Earthquake	19,846	368,820	273,218

Table showing the single events, from 1900 to 2022, that have killed more people, affected more people (injured, homeless) and produce the greatest economic impact, respectively. This helps us see the great impact that one single natural desaster can have on mankind (Research Question 1).

Findings: Greatest Types of Natural Disasters?



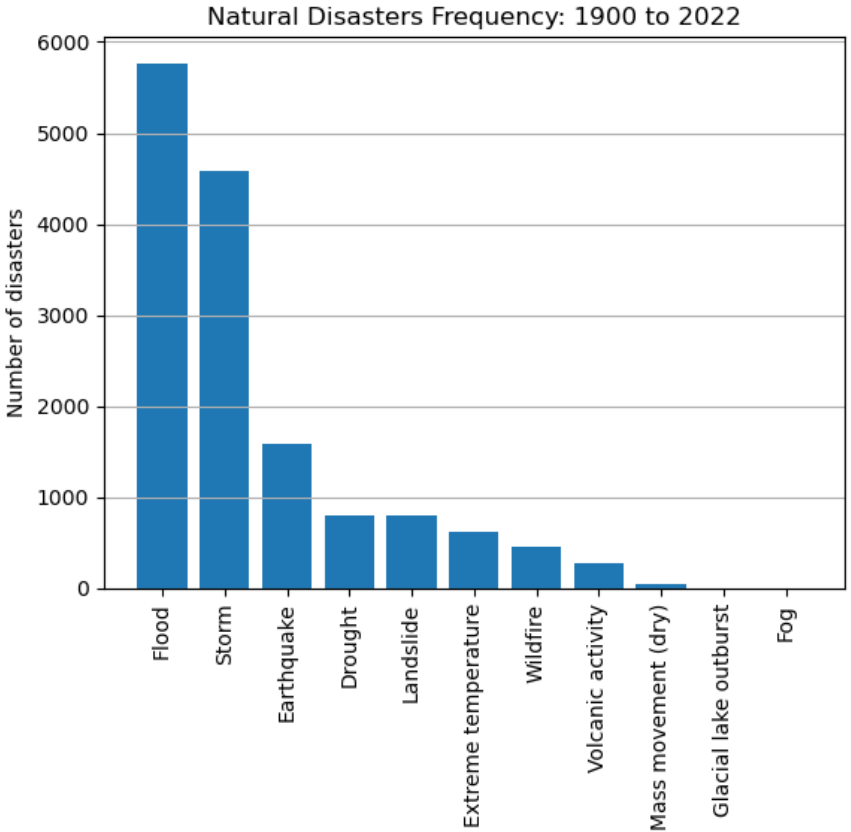
Droughts cause over 50% of all deaths from natural disasters.

Floods generates over 45% of all people affected by natural disasters.

Storms generates over 40% of all economic impact from natural disasters.

It can be said that droughts have the greatest human impact, while storms have the greatest economic impact.

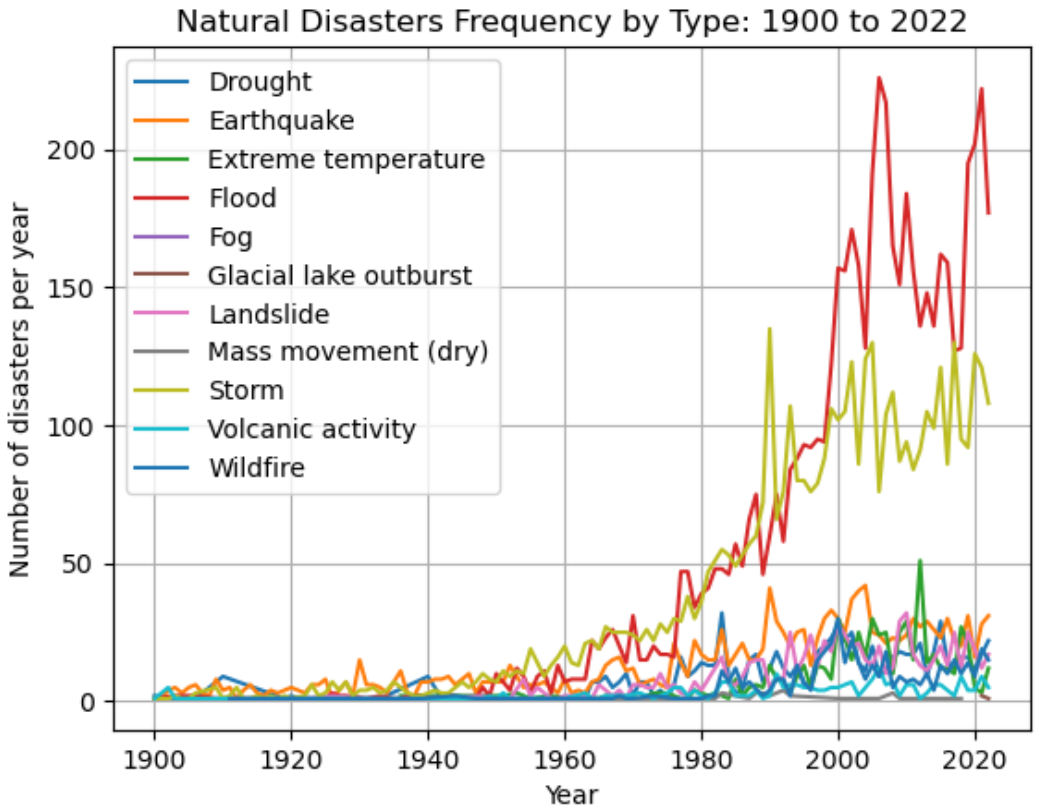
Findings: How often does Natural Disasters happen?



In this bar plot we can se how common or frequent are the different types of natural disasters.

Very quickly, it can be spotted that the most common, or frequent, natural disasters types are floods and storms, by a large margin.

Findings: How often does Natural Disasters happen?



This line graph helps us examine how often the types of natural disaster happen over time.

Especially clear since 1960, over time, floods (red line) are the most frequent disasters, followed by storms (yellow line). This was reversed on the 1990s.

This is consistent with the bar plot shown in the previous slide.

Findings: Can deaths be predicted?

It would be useful being able to predict the number of deaths caused by a type of natural disaster, by only knowing the region where it could happen and the type of disaster.

A model like this could be used by emergency response teams to prepare for possible scenarios, or for estimating deaths in semi real-time as the disaster is happening.

Regression Analysis was selected because the predicted value is a number (deaths). The categorical data was transformed to numerical data, because the regression algorithms used from scikit-learn only work with numbers.

Data was divided, 67% for training the model and 33% to test it.

Findings: Can deaths be predicted?

Target value	Deaths
Count	4 924
Mean	1 387
Standard Deviation (std)	57 389

Regression Models	Linear Regressor (LR)	Decision Tree Regressor (DTR)
Root mean square error (RSME)	57 398	57 258
Negative deaths?	Yes, 2 630 predictions were negative.	No, 0 predictions were negative.

The table on the left (orange) describes the value that is being predicted, while the table on the right shows the regression results.

Of importance is the **very high std of the target (deaths)**, characteristic of natural disaster loss data. In the light of this, even if the **RSME is large** for the two models generated, they **are in the range of 1 std**.

The **DTR is a better model** in this case, exclusively because it **does not predict negative deaths**.

Limitations

i. Disasters before the year 1900 are not consider by the dataset.

ii. The results shown are subjected to how well documented natural disasters have been over the last 100 years. It is posible that, for example, from 1900 to1950, their documentation was not as thorough as nowadays, thanks to technology.

iii. The information about the disaster magnitude was very incomplete, so it couldn’t be used. If it has been complete, it would possibly improve the accuracy of the regression analysis.

Conclusions

i. Although not so common or frequent, the deadliest disaster type is drought, but the deadliest single event was a flood (China, 1931, 3.7 million dead).

ii. Floods and storms are the most common types of disasters. This could be the reason why they are also the ones affecting most people and producing the greatest economic losses, respectively, although the single event affecting more peple was a drought (India, 2015, 330 million people) and the one of greater economic impact was an earthquake (Japan, 2015, 273 billion USD).

iii. The model generated to predict deaths caused by all types of natural disasters has large errors, related to the very high variability of the reported deaths.

Further Analysis

i. I have a hypothesis: Because, in order for a natural event to produce a disaster it has to affect people, one reason for the large increase in natural disasters since 1960 might be related to the increase of the world population. The latter in turn leads to more people living in highly populated areas like cities, and also on more dangerous areas, like river margins. So, I think that merging the dataset used in this analysis with one containing information about world population, will help answer this hypothesis.

ii. With more data, it could be posible to explore specific death prediction models for each type of disaster, to see if the error can be reduce.

Acknowledgements

Source of database:

EM-DAT The international Disaster Database, public database at https://public.emdat.be, from the Centre of Research on the Epidemiology of Disasters (CRED).

Feedback:

From my deer wife, María Echandi.

References

Database features description:

EM-DAT The international Disaster Database, database guidelines at Guidelines | EM-DAT (emdat.be), from the Centre of Research on the Epidemiology of Disasters (CRED).

Analysis:

All the analysis work was done by myself.